Київський національний університет імені Тараса Шевченка
механіко-математичний факультет
кафедра теорії ймовірностей,
статистики та актуарної математики

КУРСОВА РОБОТА

зі статистики на тему

«Прогнозування часового ряду за допомогою моделі ARMA(p, q)»

Студентки 3-го курсу
спеціальності «111 Математика»
Вербицької Олени Вікторівни

Науковий керівник:
професор, доктор фіз.-мат. наук
Шевченко Георгій Михайлович

Київ — 2021

Зміст

1 Попередній аналіз часового ряду та перетворення його на стаціонарний часовий ряд **2**

2 Пошук коефіцієнтів p та q для моделі ARMA(p, q) **5**

3 Побудова прогнозу та аналіз його якості **7**

4 Висновки **9**

5 Література **9**

1 Попередній аналіз часового ряду та перетворення його на стаціонарний часовий ряд

На **Рис.1** представлено графік дискретного часового ряду з датасету **Electricity_production.csv**. На горизонтальній осі відмічені моменти часу в місяцях, а на вертикальній – кількість виробленої електроенергії за вказаний місяць.

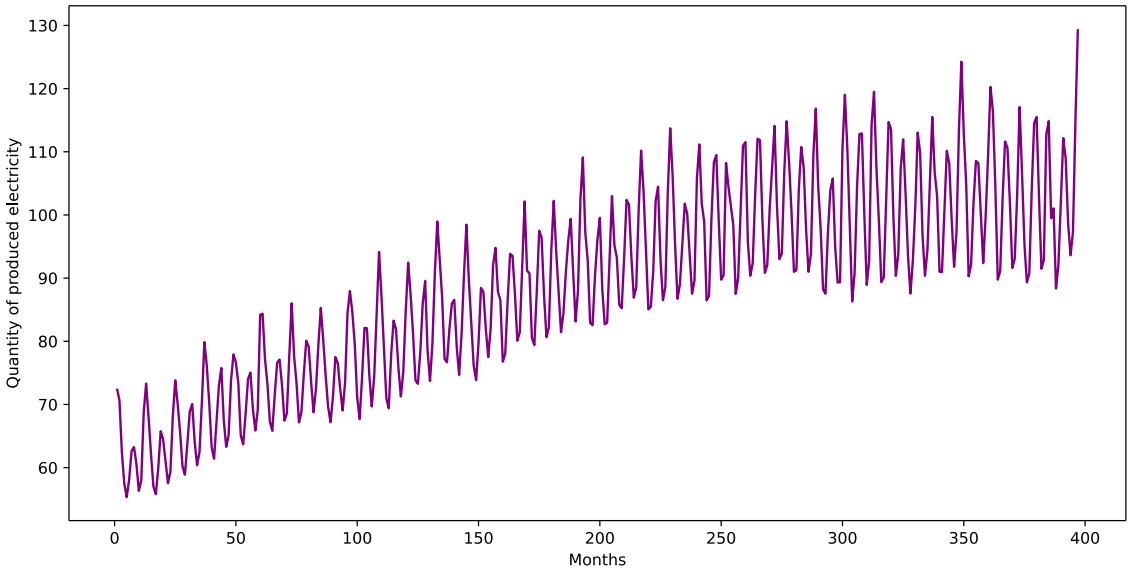

Рис. 1: Графік часового ряду y_t

В цьому графіку присутня трендова компонента, що відповідає за плавне зростання, та сезонність: кожного наступного року значення ряду змінюються приблизно так, як і минулого року. Також відмітимо зростання дисперсії з плином часу.

За теоремою Вольда кожний стаціонарний часовий ряд можна описати моделлю ARMA(p, q) з наперед заданою точністю. Тому перше, що нам необхідно зробити – це перевірити початковий часовий ряд на стаціонарність, а у випадку його нестаціонарності слід провести деякі перетворення.

Для перевірки на стаціонарність використаємо тест Дікі-Фуллера (Augmented Dickey–Fuller Test):

часовий ряд: $y \quad \Rightarrow \quad y_1,\ y_1, \ldots,\ y_t$;

початкова гіпотеза: $H_0:$ ряд нестаціонарний;

альтернатива: $H_1:$ ряд стаціонарний;

статистика: t-статистика для перевірки значущості коефіцієнтів лінійної регресії;

розподіл: розподіл Дікі-Фуллера.

За допомогою відповідної бібліотеки в Python було отримано, що початковий ряд не є стаціонарним, оскільки значення ADF Statistic не перевищує критичні значення (Critical Values) на кожному з трьох рівнів значущості $\alpha \in \{1\%,\ 5\%,\ 10\%\}$.

ADF Statistic: -2.256990
p-value: 0.186215
Critical Values:
1%: -3.448; 5%: -2.869; 10%: -2.571

Однією з причин нестаціонарності часового ряду y_t є наявність трендової компоненти, яка, звичайно, змінює значення середнього для фіксованого інтервалу часу. Щоб позбутися трендової компоненти іноді достатньо від початкового ряду перейти до ряду його попарних різниць (диференціювання часового ряду). Але оскільки часовий ряд y_t містить сезонну компоненту з періодом 12, то варто застосовувати сезонне диференціювання - перехід до попарних різниць значень часового ряду в сусідніх сезонах. Необхідне перетворення $y_1, y_2, \ldots, y_n \to y'_{13}, y'_{14}, \ldots, y'_n$ здійснюється за формулою

$$y'_t = y_t - y_{t-12}.$$

Варто зазначити, що диференціювання ряду можна робити декілька разів, поки ряд не стане стаціонарним.

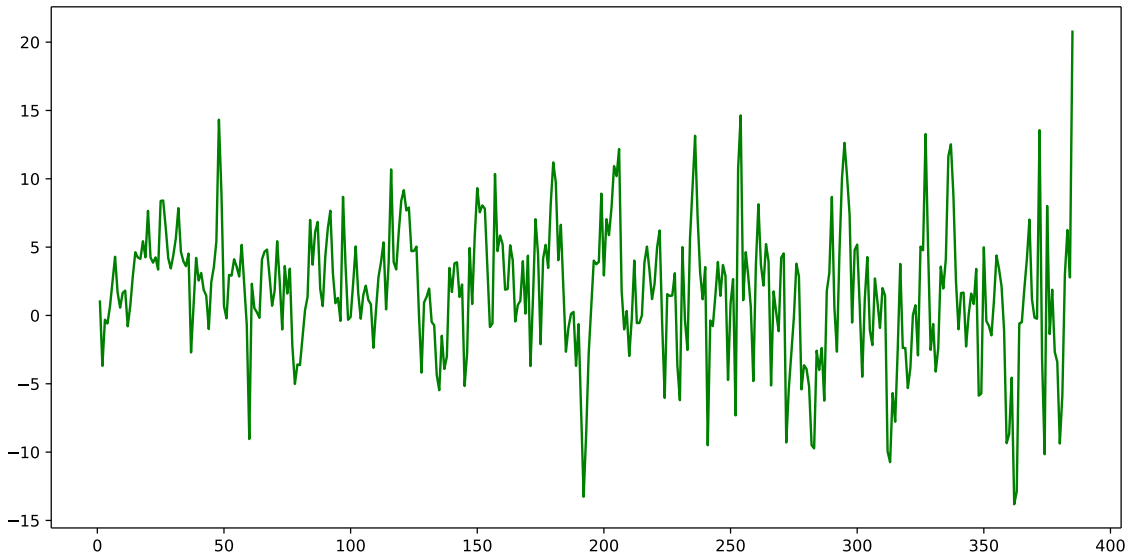

Рис. 2: Графік часового ряду y'_t

На **Рис.2** можна побачити результат проведення одного диференціювання часового ряду y_t. Перевіримо часовий ряд y'_t на стаціонарність за допомогою теста Дікі-Фуллера.

ADF Statistic: -5.673482

p-value: 0.000001

Critical Values:

1%: -3.448; 5%: -2.869; 10%: -2.571

Значення ADF Statistic є більшим за критичне значення на рівні значущості $\alpha = 1\%$. Отже, відхиляємо гіпотезу H_0 про нестаціонарність ряду на кожному з трьох рівнів значущості $\alpha \in \{1\%,\ 5\%,\ 10\%\}$. Ряд y'_t – стаціонарний.

2 Пошук коефіцієнтів p та q для моделі ARMA(p, q)

Для знаходження прогнозу за допомогою моделі ARMA(p, q) нам спочатку потрібно з'ясувати, яку кількість регресійних компонент p та яку кількість компонент рухомого середнього q необхідно розглядати, щоб ця модель мала найменші відхилення від початкового стаціонарного ряду y'_t.

За допомогою автокореляційної функції знайдемо значення q. Автокореляцією порядку k називають кореляцію між часовим рядом y'_t та його копією зсунутою на k значень, тобто

$$r_k = Corr(y_t, y_{t-k}).$$

Функцією автокореляції називають залежність значення коефіцієнту r_k від значення відповідного лагу k. **Останнє значення k при якому r_k суттєво відрізняється від нуля і буде нашим шуканим значенням**.

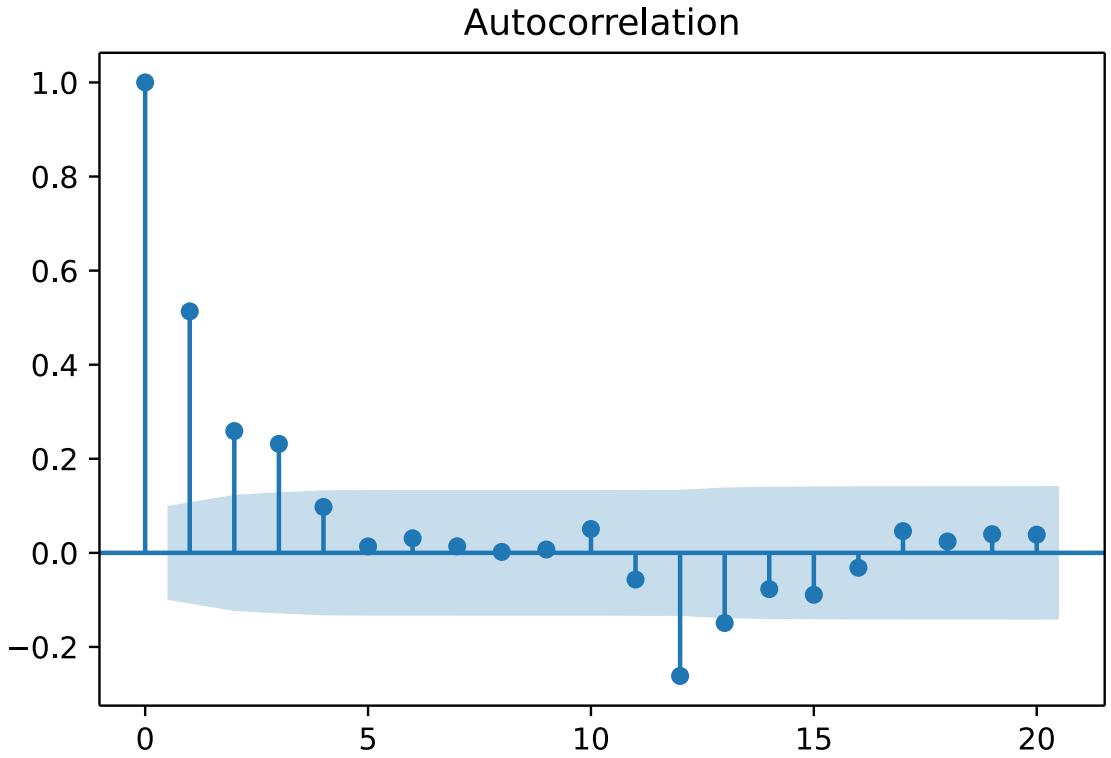

Рис. 3: Графік функції автокореляції для стаціонарного ряду y'_t

Принцип відшукання значення параметру p в моделі ARMA(p, q) будується аналогічним чином, проте його оцінка здійснюється вже не за допомогою функції автокореляції, а за допомогою функції часткової автокореляції. Часткова автокореляція порядку k визначається за формулою

$$\phi_k = Cor(y'_t - P(y'_t), y'_{t-k} - P(y'_{t-k})),$$

де $P(y'_t)$ – проекція випадкової величини y'_t на лінійну оболонку значень y'_{t-1}, y'_{t-2}, $\dots$, y'_{t-k+1}.

Загальна процедура знаходження коефіцієнту p в моделі ARMA(p, q) полягає лише в побудові функції часткової автокореляції для стаціонарного часового ряду y'_t. **Останнє значення k при якому ϕ_k суттєво відрізняється від нуля і буде нашим шуканим значенням**.

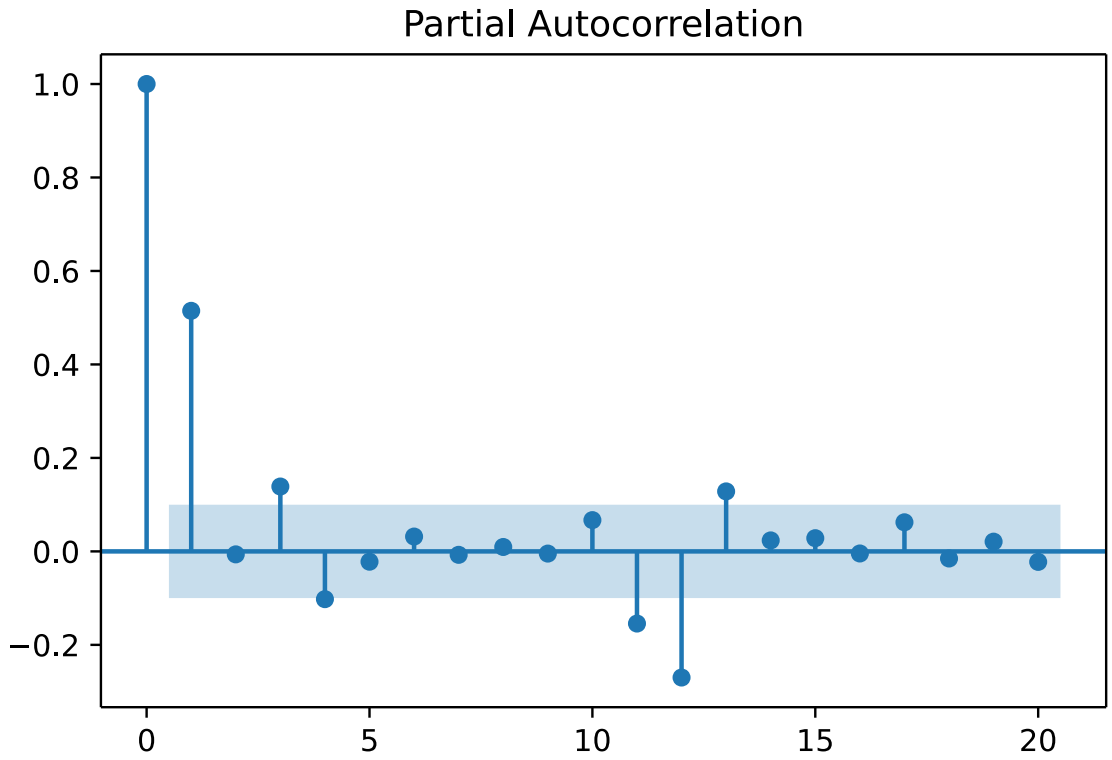

Рис. 4: Графік функції часткової автокореляції для стаціонарного ряду y'_t

Отже, оберемо параметри $p = 3$ та $q = 3$ для моделі ARMA(p, q).

3 Побудова прогнозу та аналіз його якості

Використаємо функцію ARMA(3, 3) в Python для побудови прогнозу s'_t на наступні 2 роки для стаціонарного ряду y'_t. Потім зробимо обернене диференціювання (інтегрування) ряду s'_t, щоб отримати прогноз s_t для початкового часового ряду y_t.

На **Рис.5** фіолетовий графік відповідає початковому часовому ряду y_t, а помаранчевий - спрогнозованому часовому ряду s_t.

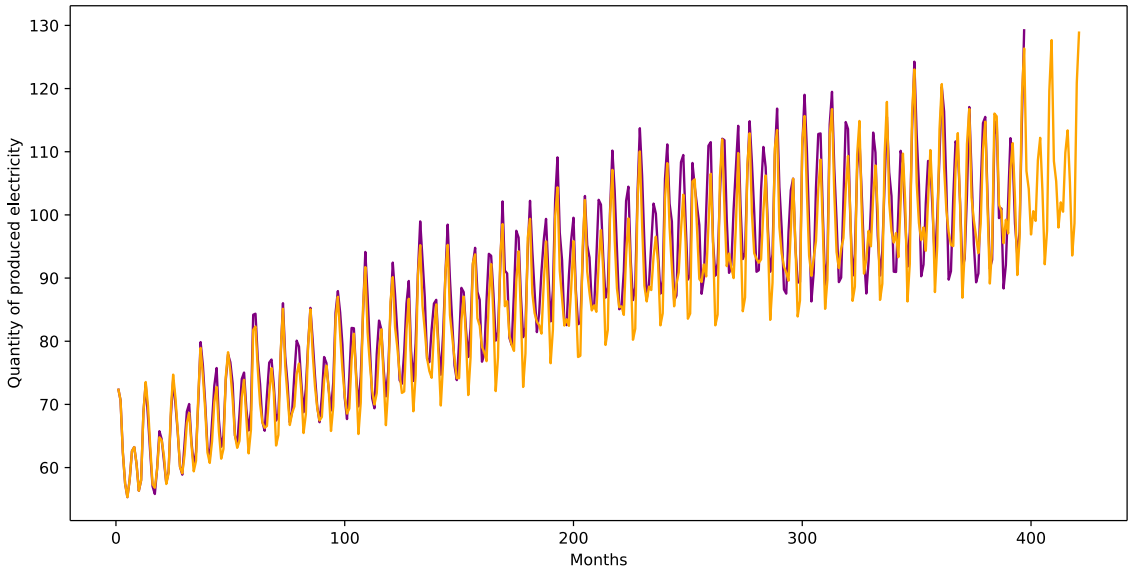

Рис. 5: Графік функції часового ряду y_t та прогнозу s_t

Щоб оцінити якість побудованої моделі, розглянемо залишки x_t – різниці реальних даних та спрогнозованих в момент часу t

$$x_t = y_t - s_t$$

Дослідимо їх розподіл на нормальність. Можна зробити це графічним методом за допомогою гістограми та графіка Q-Q (quantile-quantile plot) – графіка, що покаже залежність між теоретичними та вибірковими квантилями. Якщо ця залежність лінійна, то вибірка нормально розподілена.

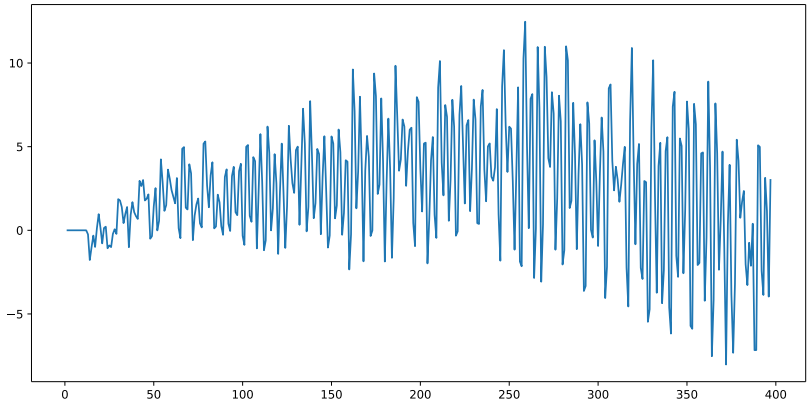

Рис. 6: Графік залишків x_t

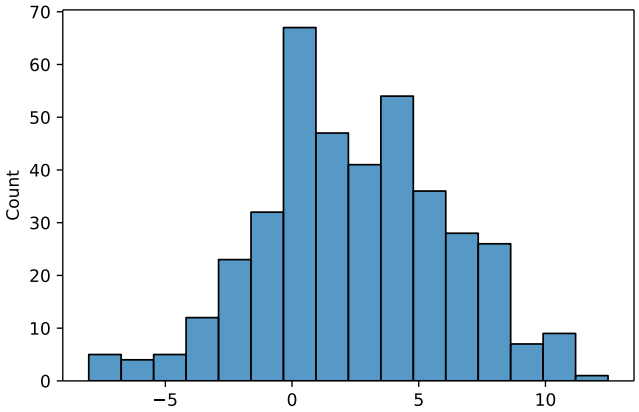

Рис. 7: Гістограма залишків x_t

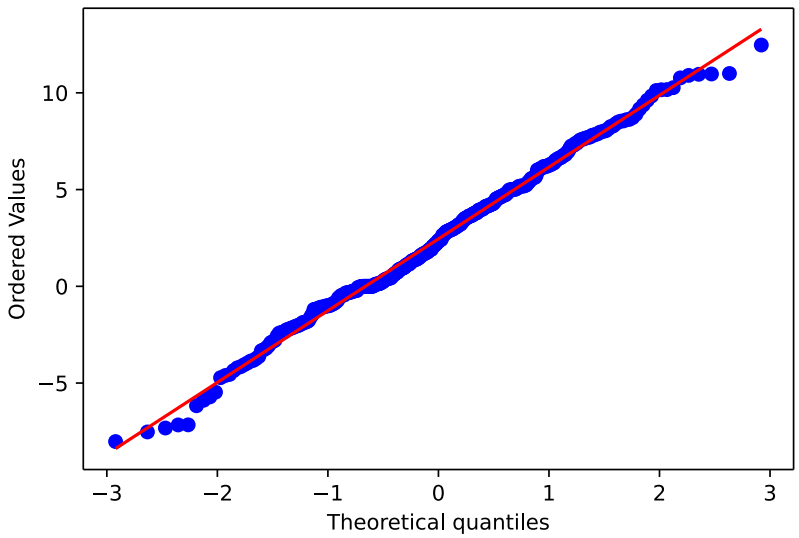

Рис. 8: Q-Q графік залишків x_t

Як бачимо, гістограма розподілу **Рис.7** майже симетрична, а точки на графіку Q-Q **Рис.8** лежать на одній прямій, тому ми можемо припустити, що розподіл залишків є нормальним. Для перевірки гіпотези H_0 (вибірка є нормально розподіленою) можна скористатися більш точним тестом на нормальність – тестом Шапіро-Уілка (Shapiro-Wilk normality test), що також є придатним і для таких невеликих вибірок, як залишки x_t. За допомогою відповідної функції в Python було знайдено $p - value = 0.229$ і підтверджено H_0. Це означає, що побудована модель добре враховує структуру початкового ряду при прогнозуванні та є досить якісною для практичного застосування.

4 Висновки

У цій роботі було проведено аналіз часового ряду з датасету **Electricity_production.csv** на можливість прогнозування його майбутніх значень за допомогою моделі ARMA(p, q). Оскільки цей ряд вдалося за скінченну кількість перетворень зробити стаціонарним, то за теоремою Вольда маємо, що цей стаціонарний ряд можна описати моделлю ARMA(p, q). За допомогою аналіза функцій автокореляції та часткової автокореляції було підібрано параметри p і q ($p = 3$, $q = 3$). За допомогою моделі ARMA(3, 3) було отримано досить гарний прогноз, якість якого було перевірено тестом Шапіро-Уілка.

5 Література

1. http://statsoft.ru/home/textbook/modules/sttimser.html
2. https://machinelearningmastery.com/arima-for-time-series-forecasting-with-python/
3. https://blog.quantinsti.com/stationarity/
4. http://www.machinelearning.ru/wiki/Критерий_Шапиро-Уилка